GSS Survey

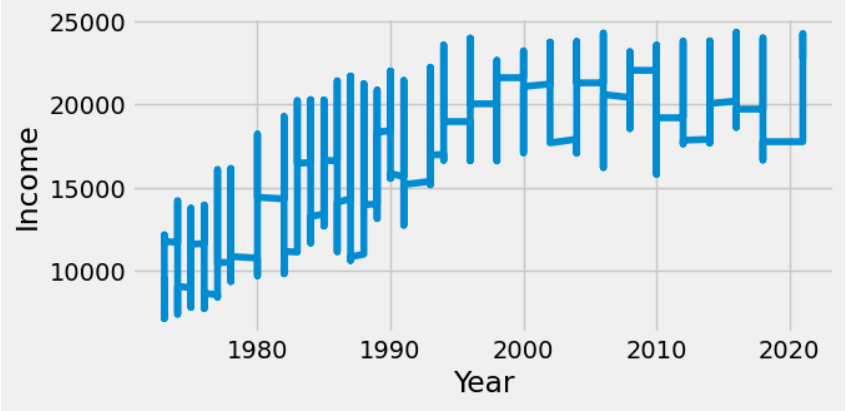
Chris Sinclair
January 19th, 2023

Task #1

(gss.py)

The first task of the GSS Survey section involved putting together a time series plot that connects the data points of “Income” and “Year”. This data – specifically the “Income” column – required extensive cleaning for all kinds of reasons so that I could convert the data type to float for calculation. The major cleaning that needed to happen was to get rid of string characters like dollar signs ($), convert string answers to integers (“$25,000 or more” became 25000), and get rid of non-answers (“Refused, “Do not Know”, “Inapplicable”, “No answer”, and “Skipped”).

Once I did that, I was able to put together a line plot that showed a rolling average of the respondents income in a per-year basis, pictured below:



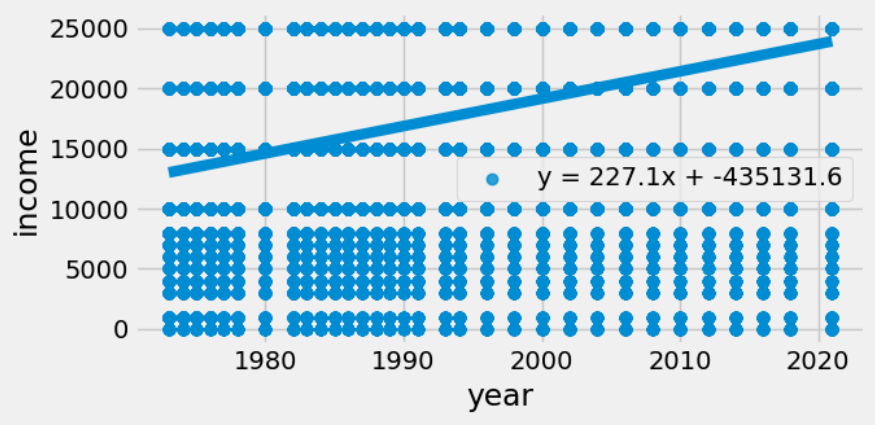
As you can see, the average income rises until the mid-2000’s and then starts to slightly dip at the tail end of the survey.

The next step was to run an ADF test to determine stationarity. Since the test statistic for the “Income” column turned out to be -11.42, which is less than the 5% critical value of -2.86, **we can conclude with 95% certainty that the “Income” variable is stationary**. My alternative hypothesis would be that income has increased over time and the null hypothesis would be that it hasn’t increased over time. Also, with our probability value calculated at 6.47e-21, we are safely under the 0.5 threshold so we can reject the null hypothesis, meaning that **the average income per year has increased over time.**

Task #2

(gss.py)

The second task required a simple linear regression. Unfortunately, I don’t think my first one went all that well in week 6, but I ran the same kind for this week 8 data and came out with the plot below:



After doing some digging into how to pull the linear function, I was able to get it printed onto the plot above as **"y = 227.1x + -435131.6"**. I manually ran some calculations to confirm that 227.1 is the correct slope, but I'm honestly not sure where the intercept of -435,131.6 is coming from. Either way, since the highest option on the survey was "$25,000 or more", and the line maxes out in 2020 right around the 25000 mark, we would need further data that stretches beyond $25,000 to confirm if the slope from this linear regression model would accurately predict data for future years, like 2024 or 2030. **With the data we've been provided as well as the limits of the maximum income option, it would be impossible to predict the accuracy of this income forecast for future years.**

Publishing Paid Me Survey

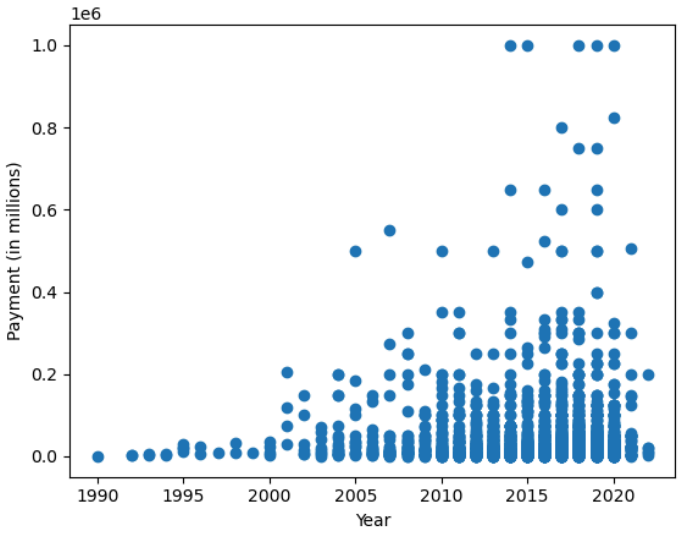
Task #1

(publishing1.py)

The first task of the Publishing Paid Me survey had us look at the two data points of **"Payment"** and **"Year"**. The hypothesis is that book advance payments have increased over time, and the probability value is less than 0.05.

I had to create a subset table for just the two relevant columns. I also cleaned the data a bit by getting rid of null values and outliers. When it comes to outliers, of the 2,200 rows, only about 10 of them consisted of payments that were over a million dollars. So I had to convert that row from objects to floats by getting rid of the "$" character and then remove any rows with payments over 1000000.0. I also removed any payments that were greater than 0, since I don't think that would count as a payment.

After that, I created a plot to visualize the data to see if there was any recognizable pattern that might support the hypothesis. The plot is pictured below:

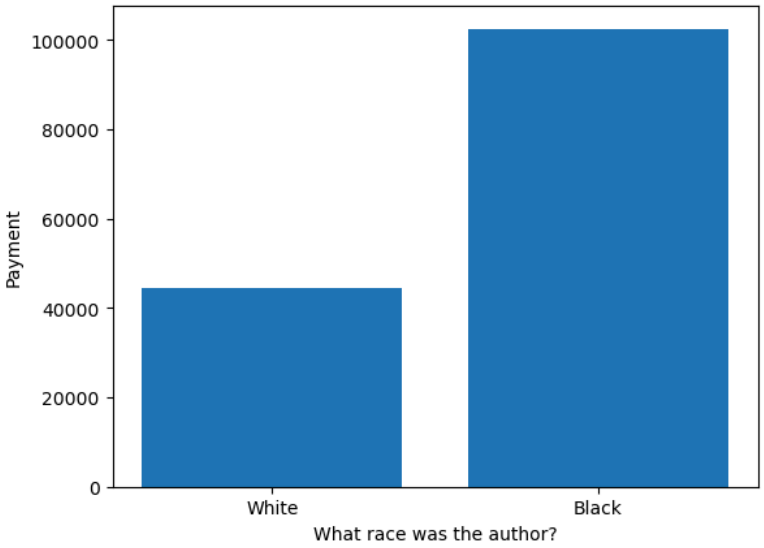


As you can see, as the years advance, so too does the payment. Visualizing the data alone doesn't really confirm the hypothesis though. We would need to crunch some numbers, so I ran a linear regression on the data and found a **probability value of 0.662,** which is significantly greater than 0.05. That means that we don't have enough evidence to soundly conclude that there is a relationship between these data points to support our hypothesis. Other factors may be at play, such as authors represented by an agency vs. authors who aren't, or the publishing company.

Task #2

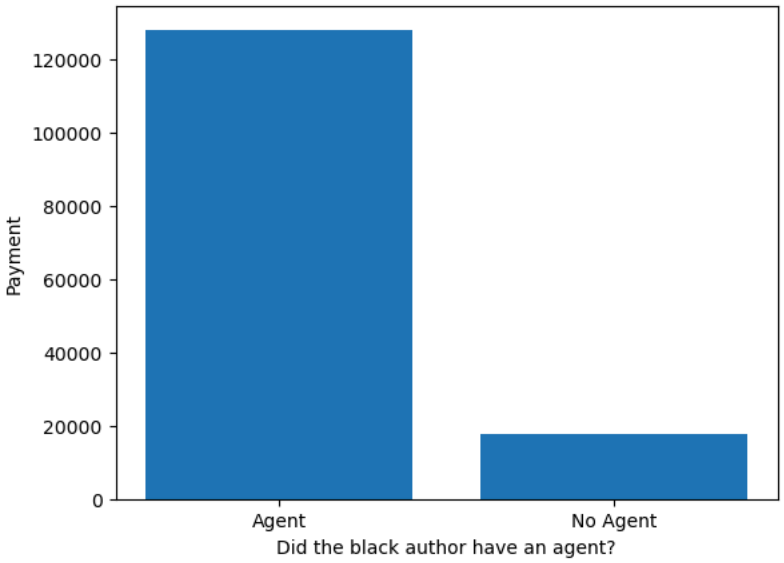
(publishing2.py)

The second task had us look at some racial differences in payment, specifically **black author payment vs. white author payment**, as well as **black authors with an agent vs. black authors without an agent.** As usual, the first step was to create our subset table and clean the data as much as possible – this means getting rid of nulls, and converting payment entries from objects to string by getting rid of the "$" character.



The mean payment for **white authors was $44,573** while the mean payment for **black authors was $102,451** – more than double that of white authors. I decided to run a linear regression on these data points, and the **correlation coefficient came out to only 0.15,** which means there is a very weak relationship between them. Also, the **probability value was 3.47,** which shows that comparing these data points is of no statistical significance. This confirms my suspicions about this chart because statistically, I've been under the assumption that white people generally find greater benefits in society across the spectrum. However, when I looked into it a bit, I found that out of the total of 1923 participants only about 100 were black authors. That small of a sample size compared to the remaining 1800 white authors shows that the data can't really be taken seriously unless you had a stronger population of black authors in the study to make for a stronger comparison.

The next part of the second task was to check if black authors with agents earned more money than black authors without agents. Since I ran this comparison in the week 2 milestone, I picked out my subset data, cleaned it up, got rid of outliers, and then calculated the mean of each to come up with the chart below:



Black authors with an agent made an average of **$128,118 per payment**, whereas black authors without an agent made only **$17,563 per payment.** Agent representation secured roughly 7.3x the payment of a black author without one. With that in mind, I did also run a linear regression on these data points to come out with a **correlation coefficient of 0.14**, which is still very weak. However, the **probability value was 0.16.** That still means there isn't statistical significance, but it is definitely a much stronger p-value than the previous tests from the tasks above.